**Analysis of Bronchoalveolar Lavage Fluid Metatranscriptome Among Patients with COVID-19 Disease**

Michael Jochum 1, Michael D. Lee 2,3, Kristen Curry 4, Viktorija Zaksas 5, Elizabeth Vitalis 6, Todd Treangen 4, Kjersti Aagaard 1, Krista Ternus 7

*1 Department of Obstetrics and Gynecology, Baylor College of Medicine and Texas Children’s Hospital, Houston, Texas 77030, USA; 2 Exobiology Branch, NASA Ames Research Center, Mountain View, California 94043, USA; 3 Blue Marble Space Institute of Science, Seattle, Washington 98104, USA; 4 Department of Computer Science, Rice University, Houston, TX, 77005, USA; 5 Center for Translational Data Science, University of Chicago, Chicago, IL, 60615, USA;* 6 Inscripta*, Inc, 5500 Central Ave STE 220, Boulder, CO, 80301,USA; 7 Signature Science, LLC, 8329 North Mopac Expressway, Austin TX*

**Abstract:**

In order to better understand the potential relationship between COVID-19 disease morbidity and microbial community dynamics / functional profiles from a hologenome standpoint, we conducted a multivariate comparison of publicly available human bronchoalveolar lavage fluid (BALF) metatranscriptomes samples amongst COVID-19 (*n*=48), community acquired pneumonia (*n*=25), and uninfected patients (*n*=32). Our overarching hypothesis was that there is a potential informative relationship between the BALF microbiome and the severity of  COVID-19 disease onset and progression. After read filtering and controlling for batch effect, the remaining SARS-CoV-2 viral and microbial reads were taxonomically classified (Kraken2), functionally characterized (SeqScreen), and analyzed for multivariable associations with COVID-19 morbidity and mortality using linear models (MaAsLin2). Among our cohort of n=105 samples, there were n=48 with COVID-19 disease, and n=20 of 48 did not survive. After controlling for differences in publication and study design, we observed significantly unique taxonomic and functional changes to the hologenome associated with COVID-19 disease and death.  Specifically, functional profiles within microbial communities predicting gene ontology classifications and proteins were significantly associated with disease severity, morbidity and death. Collectively, while this data does not speak to causality nor directionality of the association, it does demonstrate a significant relationship between the human microbiome and COVID-19 morbidity and mortality, rendering testable hypotheses that warrant further investigation.

**Introduction**

To better understand the potential relationship between COVID--19 morbidity and mortality and the human-microbiome hologenome, we conducted an analysis using human bronchoalveolar lavage fluid (BALF) metatranscriptome sample sequences sourced from 8 different public data repositories. These data arose from BALF specimens from individual subjects cohorted by one of three classifiers: (1) uninfected controls, (2) community acquired pneumonia or CAP patients, or (3) COVID-19 patients with moderate to severe disease, including death (Table 1). The objectives of the current study were to compare the BALF metatranscriptome amongst and between each of the three disease cohort classifiers, and identify significantly associated taxonomic changes in microbial derived community dynamics / and functional changes derived from gene ontologies. Our overarching testable hypothesis was that there is a potential informative and discernably significant relationship between the BALF microbiome and the severity of COVID-19 disease, including death.

**Methods**

*Metadata sources.* Supplemental Tables 1 and 2 describe the publicly available Illumina reads that were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) or the China National Center for Bioinformation (CNCB) National Genomics Data Center (NGDC), including the citations of the original publications where the clinical information was obtained for downstream analysis (1-8). Supplemental Table 3 lists the specific commands and additional details used for downstream analysis of the reads.

*Analyses*. After the raw reads were downloaded from their sources (Supplemental Tables 1-3), the quality of the reads was assessed before and after trimming with FastQC (9),and quality control was performed on the downloaded sequence reads with Trimmomatic (10). To control for different sequencing approaches by dataset (e.g.,t datasets being paired or single-end reads), all paired-end reads were converted to single-end by merging reads with flash (11), and then all merged and unmerged forward reads were combined into one file after being processed with Trimmomatic. Human and PhiX reads were filtered out with a custom Kraken2 database (12) and low complexity sequences were removed with fastp (13). Taxonomic analysis was subsequently performed on Kraken2 (12). The processed fastq datasets with human and PhiX reads removed were converted to fasta files and analyzed with SeqScreen (15) to obtain a list of leaf node molecular function and biological process Gene Ontology (GO) terms and proteins present within each of the samples. The CoV-IRT-Micro conda package (16) was used to propagate parent GO terms, parse GO terms by kingdom-level domains, and summarize Kraken2 taxonomic results and SeqScreen-reported protein identifiers.

Parent-propagated GO term counts for all domains other than eukaroytes were imported into a working phyloseq (17) object fike, alongside collected and curated clinical metadata using R 4.03 (18). Samples types of “unknown”, “sick”, and were pruned from subsequent analysis. Taxonomic classifications were decontaminated against negative controls when negative congrols were present using the library decontam to identify and remove potential contaminating organisms (14). Samples from Michalovich *et. al* (6) and samples from Shen et al. (5) that were viral enriched (PRJNA605907) were also pruned from subsequent analysis due to observed batch effects (Supplementary File 1a). After read filtering and batch effect sample removal, sample cohorts consisted of n=29 uninfected suubjects samples, n=25 CAP subjects samples, and n=32 COVID-19 subjects samples were available for comparison (total, n=86 inidvidual subjects BALF samples). Amongst the COVID-19 cohort with known survival outcomes at the time of the index study publication, n=10 were known deceased, n=15 were alive; n=7 of 32 COVID-19 subjects in this metaanalysis did not have published outcomes. GO term abundances from the remaining subjects specimens were then compositionally transformed and compared by case type (min abundance=0.01, min prevalence=0.1 normalization=CLR, and outcome (COVID-19 only) via MaAsLin2 (19) (Supplementary File 1b), controlling for random effects of publication and sample name, max significance cutoff of q < 0.05 with Benjamini-Hochberg multiple test correction (20). Additionally, GO term counts were square root transformed and subjected to community typing with Dirichlet Multinomial Mixtures (21) (Supplementary File 1b). Statistically significant GO terms were thereafter ordered by parental lineage and visualized alongside consensus DMM clusters and metadata columns publication, case, and outcome using the bioinformatic software packages pheatmap (v1.0.12) (22). Heat tress taxonomic comparisons were visualized using the bioinformatic software packages and metacoder (<https://cran.r-project.org/web/packages/metacoder/citation.html#:~:text=metacoder%20citation%20info,doi%3A%2010.1371%2Fjournal>) .

**Results**

*Comparison between disease classifiers*(uninfected controls, or patients with community acquired pneumonia (CAP) or COVID-19 disease)

After controlling for random effects of publication and patient, results from the MaAsLin2 comparison across individual subjects cohorted by one of three classifiers: (1) uninfected controls, (2) CAP patients, or (3) COVID-19 patients with moderate to severe disease, including death (Table 1)reveled 35 out of 13534 GO Terms associated with patients with COVID-19 when compared to patients with CAP but not COVID-19 or uninfected control subjects (Table 2; Table 3). Significant GO terms were comprised of 6 Depth 1 Parents involving catalytic activity [GO:0003824], binding, metabolic [GO:0008152] and cellular processes [GO:0009987], biological regulation [GO:0065007], and interspecies interaction between organisms [GO:0044419].

Significant Terms of interest associated with metatranscriptomes of BALF specimens from COVID-19 patients included hydrolase [GO:0016787] / transferase [GO:0016740] activity transferring phosphorus [GO:0016772], nucleotidyltransferase activity [GO:0016779], and ion binding [GO:0043167] ](Figure 1). Results from the Dirichlet Multinomial Mixtures clustering analysis using all 13,534 Gene ontologies counts resulted in a best model fit using 3 distinct clusters that were significantly associated with each subject or patient cohort p<0.0001] (Figure 1).

Taxonomic analysis revealed a statistically significant decrease in log2 median ration of several microbial species belonging to the genus *Sphingomonas* among BALF specimens from COVID-19 patients when compared to both the uninfected (p<0.0001, q <0.001) and CAP cohorts (p<0.005, q <0.05) (Figure 2). Interestingly, analysis of the GO Terms derived from *Sphingomonas* proteins in BALF specimens derived frompatiens with COVID-19, irrespective of disease outcomes,were

hydrogen peroxide catabolic process [GO:0042744]; response to oxidative stress [GO:0006979]

catalase activity [GO:0004096]; heme binding [GO:0020037]; and metal ion binding [GO:0046872] (Supplementary Files 3-X Krista’s protein tables).

*Metatranscriptomic comparison of BALF specimens from COVID-19 subjects stratified by disease severity or death .*

A stratified analysis amongst COVID-19 subjects specimens with known survival outcomes (of 32 subjects, n=10 were known deceased, n=15 were known to be alive) via MaAsLin2. We observed 25 unique GO terms which were significantly increased in their association with death from COVID-19 diease, with Depth 1 parents Cellular and Metabolic processes [GO:0008152; GO:0009987], Catalytic Activity [GO:0003824 ], and binding [GO:0005488], with notable functional profiles associated phosphate / phosphorylation [GO:0016310], metal ion binding (Mg, Zn, etc.) [GO:0046914;GO:0000287;GO:0008270], RNA binding [GO:0003723], and lytic activity (hydrolase, endopeptidase, oxidoreductase, etc.) [GO:0016491;GO:0016817; GO:0140098] (Table 4) (Figure 3). Features of particular interest associated with morbidity in the Depth 1 parent Metabolic and cellular processes include decreases in carbohydrate metabolic processes, Increases in RNA metabolic processes and RNA phosphodiester bond hydrolysis, decreases in phosphorylation, and increases in nucleobase containing compound biosynthetic processes.

Gene ontology comparisons amongst BALF specimens from patients with COVID-19 who suffered mortal disease demonstrated significant decreases in oxidoreductase activity, increases in catalytic activity acting on RNA, and endopeptidase activity. Lastly with respect to the Go Terms belonging to the Depth 1 parent of binding, we observed decreases in organic cyclic compound binding and increases in RNA binding transition metal ion binding, magnesium ion binding, and zinc ion binding in association with mortal COVID-19 disease (Supplementary Files 3-X Krista’s protein tables).

Results from the Taxonomic comparison analysis revealed a statistically significant increase in log2 median ratio of the family *Comanomonadacea*, belonging to the genus *Variovorax* and decreases in the family *Bacteriodales* when comparing the deceased to the survive (p<0.0001, q <0.001) (Figure 4) (Table 5).

**Discussion**

* *What are these go terms / what are they telling us*
* *Who else has found similar stuff*
* *What are these taxa telling us*

*Who else has found similar stuff*

*Functional*

We observed significantly unique discriminant taxonomic and functionalfeatures in the brochoalveolar lavage metatranscriptomes in association with COVID-19 disease and predicted its mortality. Functionally annotated Gene ontologies of interest included associated with: Phosphate / phosphorylation, metal ion binding (Mg, Zn, etc), nucleotide terms (DNA/RNA), Lytic activity (hydrolase, endopeptidase). Of note, due to limitations in the depth of clinical metadata by subject, we could not distinguish between COVID-19 pathophysiology or associated medical comorbidities, treatments, nor interventions. However, because of the time interval in which subject and patient specimens were recruited to their respective index studies (xx to YY, 2020), COVID-19-specific interventions and treatments had yet to be introduced and thus comparisons between CAP and COVID-19 subject specimens would be less likely to be related to disease-focused therapy.

The presence of SARS-CoV-2 viral proteins are driving the following GO term associations in deceased vs. survived: transition metal ion binding (GO:0046914), zinc ion binding (GO:0008270), organic cyclic compound binding (GO:0097159), and RNA binding (GO:0003723). SARS-CoV-2 proteins are also driving the associations of the following GO terms in COVID-19 vs. uninfected: modulation by symbiont of host cellular process (GO.0044068), modulation by virus of host cellular process (GO.0019054), modulation by virus of host process (GO.0019048), modulation of process of other organism involved in symbiotic interaction (GO.0051817), modulation by symbiont of host process (GO.0044003), interaction with host (GO.0016032), viral process (GO.0051701), interspecies interaction between organisms (GO.0044419), modulation by symbiont of host cellular process (GO.0044068), and modulation by virus of host cellular process (GO.0019054).

*Taxonomic comparisons*

Distinct taxonomic features of BALF specimens from COVID-19 patients with severe or mortal disease mortality include increases in log2 median ratios of genera *Sphingomonas* and *Variovorax* belonging to the *Sphingomonadacae* and *Comonomonadacea* families and decreases in the class *Bacteroidia* belonging to the order *Bacteroidiales*. These finding support previous reports regarding an association with *Sphingomonas* **[CITE ME]**, which is a not uncommon opportunistic pathogen found in nosocomial infections.

**[Define and tie the taxa to the GO terms throught the use of the Uniprot things here]**

Proteins derived from *Sphingomonas* contributed to the Sig. GO terms of interest hydrogen peroxide catabolic process [GO:0042744]; response to oxidative stress [GO:0006979]

catalase activity [GO:0004096]; heme binding [GO:0020037]; and metal ion binding [GO:0046872] amongst the COVID-19 cohort. The catalase protein decomposes hydrogen peroxide into water and oxygen; serves to protect cells from the toxic effects of hydrogen peroxide, which may suggest that *Sphingomonas* spp. responds to COVID-19 conditions in the patient by expressing genes that help it to survive well in environments undergoing great amounts of oxidative stress.

[PUT OTHER STUFF ABOUT VARIOVORAX and BACTEROIDIA HERE]

[DEFINED GO TERMS]

* Nucleobase containing compound biosynthetic process is defined as: “The chemical reactions and pathways resulting in the formation of nucleobases, nucleosides, nucleotides and nucleic acids.”
* RNA phosphodiester bond hydrolysis exonucleolytic is defined as “The chemical reactions and pathways involving the hydrolysis of terminal 3',5'-phosphodiester bonds in one or two strands of ribonucleotides.”
* Catalytic activity acting on RNA is defined as Catalytic activity that acts to modify RNA., Oxidoreductase activity is defined as “*Catalysis of an oxidation-reduction or redox reaction, a reversible chemical reaction in which the oxidation state of an atom or atoms within a molecule is altered via one substrate acting as an electron donor and becoming oxidized, while the other acts as an electron acceptor and becomes reduced.”*
* The Terminal GO Term endopeptidase activity is defined as “*Catalysis of the hydrolysis of internal, alpha-peptide bonds in a polypeptide chain.”*
* RNA binding is defined as *Interacting selectively and non-covalently with an RNA molecule or a portion thereof.* *Ion binding is defined as Interacting selectively and non-covalently with magnesium (Mg) and (Zn) ions.*

Collectively, while this data does cannot speak to causality or directionality of the association, it does demonstrate a significant relationship between the human microbiome and severity of COVID-19, rendering further testable hypotheses that warrant further investigation.

**Acknowledgments**

We would like to thank the COVIRT microbial subgroup team members and give special acknowledgment to John Fonner and the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported.

**References**

1. Chen L, Liu W, Zhang Q, Xu K, Ye G, Wu W, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg Microbes Infect. 2020;9: 313–319.

2. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579: 265–269.

3. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579: 270–273.

4. Xiong Y, Liu Y, Cao L, Wang D, Guo M, Jiang A, et al. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. Emerg Microbes Infect. 2020;9: 761–770.

5. Shen Z, Xiao Y, Kang L, Ma W, Shi L, Zhang L, et al. Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients with Coronavirus Disease 2019. Clinical Infectious Diseases. 2020. doi:10.1093/cid/ciaa203

6. Michalovich D, Rodriguez-Perez N, Smolinska S, Pirozynski M, Mayhew D, Uddin S, et al. Obesity and disease severity magnify disturbed microbiome-immune interactions in asthma patients. Nat Commun. 2019;10: 5711.

7. Huang W, Yin C, Wang G, Rosenblum J, Krishnan S, Dimitrova N, et al. Optimizing a Metatranscriptomic Next-Generation Sequencing Protocol for Bronchoalveolar Lavage Diagnostics. J Mol Diagn. 2019;21: 251–261.

8. Ren L, Zhang R, Rao J, Xiao Y, Zhang Z, Yang B, et al. Transcriptionally Active Lung Microbiome and Its Association with Bacterial Biomass and Host Inflammatory Status. mSystems. 2018;3. doi:10.1128/mSystems.00199-18

9. FastQC

10. Trimmomatic

11. flash

12. Kraken2

13. fastp

14. decontam

15. SeqScreen

16. CoV-IRT-Micro conda package

17. McMurdie PJ, Holmes S: **phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data**. *PLOS ONE* 2013, **8**(4):e61217.

18. R 4.03

19. MaAsLin2

20. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful**

**approach to multiple testing**. *Journal of the Royal statistical society: series B (Methodological)*

1995, **57**(1):289-300.

21. Holmes I, Harris K, Quince C: **Dirichlet Multinomial Mixtures: Generative Models for**

**Microbial Metagenomics**. *PLOS ONE* 2012, **7**(2):e30126.

22. Kolde R: **Pheatmap: pretty heatmaps**. *R package version* 2012, **1**(2).

Table 1. Overview of Meta-analysis dataset Clinical Characteristics (*n*=86)

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | Uninfected | Community acquired pneumonia | COVID-19 |
| **case** | 29 (33.72%) | 25 (29.07%) | 32(37.21%) |
| **Outcome  (COVID – 19 only)** |  |  |  |
| Deceased | - | - | 10 (31.25%) |
| Survived | - | - | 15 (46.87%) |
| Unspecified | - | - | 7 (21.88%) |
| **Sex** |  |  |  |
| female | 4 (18.18%) | 8 (36.36%) | 10 (45.45%) |
| male | 5 (13.15%) | 11 (28.94%) | 22 (57.89%) |
| unspecified | 20 (76.92%) | 6(23.07%) | 0 (0%) |
| **Reads** |  |  |  |
| paired | 29 (37.18%) | 25 (32.05%) | 24 (30.77%) |
| single | 0 (0%) | 0 (0%) | 8 (100%) |
| unspecified | 0 (0%) | 0 (0%) | 0 (0%) |
| **Publication** |  |  |  |
| Chen | 0 (0%) | 0 (0%) | 2 (100%) |
| Ren | 9 (100%) | 0 (0%) | 0 (0%) |
| Shen | 20 (32.79%) | 25 (40.98%) | 16 (40.98%) |
| Wu | 0 (0%) | 0 (0%) | 1 (100%) |
| Xiong | 0 (0%) | 0 (0%) | 4 (100%) |
| Zhou | 0 (0%) | 0 (0%) | 9 (100%) |
| **Numeric variables (**mean ± SD) |  |  |  |
| Age | 53.2 ± 13.3 (n=9) | 51.2 ± 19.8 (n=17) | 47.3 ± 11.5 (n=32) |
| Temp. °C | - | 38.4 ± 0.91 (n=15) | 38.4 ± 0.715 (n=8) |
| days after onset | - | 9.07 ± 3.17 (n=14) | 12.05 ± 6.5 (n=41) |

Table 2. MaAsLin2 derived significant Gene Ontologies associated with COVID-19 (n=32) when compared to the Community acquired Pneumonia (n=25) cohort. Write something here about transformation normalization controlling for random effect and benjamini Hochberg correction.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| feature | namespace | value (vs COVID-19) | coef | stderr | N | N.not.0 | pval | qval |
| hydrolase activity | GO:0016787 | Community Acquired Pneumonia | -0.008 | 0.001 | 86.000 | 69.000 | 0.000 | 0.000 |
| cellular process | GO:0009987 | Community Acquired Pneumonia | 0.010 | 0.002 | 86.000 | 74.000 | 0.000 | 0.000 |
| transferase activity | GO:0016740 | Community Acquired Pneumonia | 0.006 | 0.002 | 86.000 | 58.000 | 0.000 | 0.001 |
| modulation by virus of host cellular process | GO:0019054 | Community Acquired Pneumonia | -0.004 | 0.001 | 86.000 | 11.000 | 0.000 | 0.001 |
| modulation by symbiont of host cellular process | GO:0044068 | Community Acquired Pneumonia | -0.004 | 0.001 | 86.000 | 11.000 | 0.000 | 0.001 |
| biosynthetic process | GO:0009058 | Community Acquired Pneumonia | 0.005 | 0.001 | 86.000 | 23.000 | 0.000 | 0.001 |
| cellular macromolecule metabolic process | GO:0044260 | Community Acquired Pneumonia | 0.002 | 0.001 | 86.000 | 4.000 | 0.001 | 0.003 |
| organic substance biosynthetic process | GO1901576 | Community Acquired Pneumonia | 0.003 | 0.001 | 86.000 | 12.000 | 0.001 | 0.004 |
| cellular biosynthetic process | GO:0044249 | Community Acquired Pneumonia | 0.003 | 0.001 | 86.000 | 12.000 | 0.001 | 0.004 |
| transferase activity transferring phosphorus containing groups | GO:0016772 | Community Acquired Pneumonia | 0.005 | 0.001 | 86.000 | 14.000 | 0.001 | 0.004 |
| cellular metabolic process | GO:0044237 | Community Acquired Pneumonia | 0.004 | 0.001 | 86.000 | 72.000 | 0.005 | 0.012 |
| modulation by virus of host process | GO:0019048 | Community Acquired Pneumonia | -0.004 | 0.001 | 86.000 | 18.000 | 0.005 | 0.012 |
| nucleotidyltransferase activity | GO:0016779 | Community Acquired Pneumonia | 0.003 | 0.001 | 86.000 | 7.000 | 0.006 | 0.012 |
| metabolic process | GO:0008152 | Community Acquired Pneumonia | 0.005 | 0.002 | 86.000 | 76.000 | 0.006 | 0.013 |
| organonitrogen compound metabolic process | GO:1901564 | Community Acquired Pneumonia | 0.002 | 0.001 | 86.000 | 5.000 | 0.009 | 0.018 |
| modulation by symbiont of host process | GO:0044003 | Community Acquired Pneumonia | -0.003 | 0.001 | 86.000 | 18.000 | 0.018 | 0.032 |
| modulation of process of other organism involved in symbiotic interaction | GO:0051817 | Community Acquired Pneumonia | -0.003 | 0.001 | 86.000 | 18.000 | 0.018 | 0.032 |
| modulation of process of other organism | GO:0035821 | Community Acquired Pneumonia | -0.003 | 0.001 | 86.000 | 18.000 | 0.018 | 0.032 |
| organic substance metabolic process | GO:0071704 | Community Acquired Pneumonia | 0.003 | 0.001 | 86.000 | 76.000 | 0.026 | 0.045 |
| nucleic acid metabolic process | GO:0090304 | Community Acquired Pneumonia | -0.002 | 0.001 | 86.000 | 17.000 | 0.029 | 0.048 |

Table 3. MaAsLin2 derived significant Gene ontologies associated with COVID-19 (n=32) when compared to the Uninfected (n=29) cohort. Write something here about transformation normalization controlling for random effect and benjamini Hochberg correction.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| feature | namespace | value (vs COVID-19) | coef | stderr | N | N.not.0 | pval | qval |
| cellular process | GO:0009987 | Uninfected | 0.016 | 0.002 | 86.000 | 74.000 | 0.000 | 0.000 |
| metabolic process | GO:0008152 | Uninfected | 0.013 | 0.002 | 86.000 | 76.000 | 0.000 | 0.000 |
| modulation by symbiont of host cellular process | GO:0044068 | Uninfected | -0.007 | 0.001 | 86.000 | 11.000 | 0.000 | 0.000 |
| modulation by virus of host cellular process | GO:0019054 | Uninfected | -0.007 | 0.001 | 86.000 | 11.000 | 0.000 | 0.000 |
| modulation by virus of host process | GO:0019048 | Uninfected | -0.008 | 0.001 | 86.000 | 18.000 | 0.000 | 0.000 |
| organic substance metabolic process | GO:0071704 | Uninfected | 0.008 | 0.001 | 86.000 | 76.000 | 0.000 | 0.000 |
| cellular macromolecule metabolic process | GO:0044260 | Uninfected | 0.004 | 0.001 | 86.000 | 4.000 | 0.000 | 0.000 |
| cellular metabolic process | GO:0044237 | Uninfected | 0.009 | 0.001 | 86.000 | 72.000 | 0.000 | 0.000 |
| modulation by symbiont of host process | GO:0044003 | Uninfected | -0.008 | 0.001 | 86.000 | 18.000 | 0.000 | 0.000 |
| modulation of process of other organism | GO:0035821 | Uninfected | -0.008 | 0.001 | 86.000 | 18.000 | 0.000 | 0.000 |
| modulation of process of other organism involved in symbiotic interaction | GO:0051817 | Uninfected | -0.008 | 0.001 | 86.000 | 18.000 | 0.000 | 0.000 |
| hydrolase activity | GO:0016787 | Uninfected | -0.008 | 0.001 | 86.000 | 69.000 | 0.000 | 0.000 |
| interaction with host | GO:0051701 | Uninfected | -0.009 | 0.002 | 86.000 | 20.000 | 0.000 | 0.000 |
| viral process | GO:0016032 | Uninfected | -0.013 | 0.002 | 86.000 | 31.000 | 0.000 | 0.000 |
| transferase activity | GO:0016740 | Uninfected | 0.009 | 0.002 | 86.000 | 58.000 | 0.000 | 0.000 |
| primary metabolic process | GO:0044238 | Uninfected | 0.006 | 0.001 | 86.000 | 74.000 | 0.000 | 0.000 |
| symbiotic process | GO:0044403 | Uninfected | -0.014 | 0.002 | 86.000 | 31.000 | 0.000 | 0.000 |
| interspecies interaction between organisms | GO:0044419 | Uninfected | -0.014 | 0.002 | 86.000 | 31.000 | 0.000 | 0.000 |
| macromolecule metabolic process | GO:0043170 | Uninfected | 0.004 | 0.001 | 86.000 | 66.000 | 0.000 | 0.000 |
| organonitrogen compound metabolic process | GO:1901564 | Uninfected | 0.004 | 0.001 | 86.000 | 5.000 | 0.000 | 0.000 |
| binding | GO:0005488 | Uninfected | 0.004 | 0.001 | 86.000 | 81.000 | 0.000 | 0.001 |
| nitrogen compound metabolic process | GO:0006807 | Uninfected | 0.004 | 0.001 | 86.000 | 70.000 | 0.000 | 0.001 |
| biosynthetic process | GO:0009058 | Uninfected | 0.004 | 0.001 | 86.000 | 23.000 | 0.005 | 0.012 |
| ion binding | GO:0043167 | Uninfected | 0.002 | 0.001 | 86.000 | 8.000 | 0.006 | 0.012 |
| regulation of biological process | GO:0050789 | Uninfected | -0.003 | 0.001 | 86.000 | 15.000 | 0.010 | 0.020 |
| cellular nitrogen compound metabolic process | GO:0034641 | Uninfected | 0.002 | 0.001 | 86.000 | 53.000 | 0.011 | 0.021 |
| transferase activity transferring phosphorus containing groups | GO:0016772 | Uninfected | 0.004 | 0.001 | 86.000 | 14.000 | 0.014 | 0.027 |
| catalytic activity | GO:0003824 | Uninfected | 0.009 | 0.004 | 86.000 | 86.000 | 0.023 | 0.041 |
| RNA metabolic process | GO:0016070 | Uninfected | 0.002 | 0.001 | 86.000 | 6.000 | 0.028 | 0.048 |
| regulation of cellular process | GO:0050794 | Uninfected | -0.002 | 0.001 | 86.000 | 12.000 | 0.030 | 0.050 |

Table 4. MaAsLin2 derived significant Gene ontologies associated with COVID-19 disease outcome (deceased vs. survived). *Write something here about transformation normilzation controlling for random effect and Benjamini-Hochberg multiple test correction*.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| name | ontology | namespace | depth | coef | stderr | pval | qval | N | N.not.zero |
| pyrimidine-containing compound metabolic process | biological\_process | GO:0072527 | 4 | -4.815 | 0.867 | <0.001 | <0.001 | 25 | 12 |
| nucleobase-containing compound biosynthetic process | biological\_process | GO:0034654 | 5 | -0.630 | 0.117 | <0.001 | <0.001 | 25 | 25 |
| transition metal ion binding | molecular\_function | GO:0046914 | 5 | -0.545 | 0.106 | <0.001 | <0.001 | 25 | 25 |
| aromatic compound biosynthetic process | biological\_process | GO:0019438 | 4 | -0.478 | 0.116 | <0.001 | 0.004 | 25 | 25 |
| heterocycle biosynthetic process | biological\_process | GO:0018130 | 4 | -0.393 | 0.100 | <0.001 | 0.007 | 25 | 25 |
| macromolecule biosynthetic process | biological\_process | GO:0009059 | 4 | 0.382 | 0.103 | <0.001 | 0.015 | 25 | 25 |
| RNA metabolic process | biological\_process | GO:0016070 | 6 | -0.310 | 0.086 | <0.001 | 0.018 | 25 | 25 |
| RNA phosphodiester bond hydrolysis | biological\_process | GO:0090501 | 7 | -1.412 | 0.402 | <0.001 | 0.024 | 25 | 17 |
| magnesium ion binding | molecular\_function | GO:0000287 | 5 | -2.336 | 0.709 | 0.001 | 0.036 | 25 | 11 |
| RNA binding | molecular\_function | GO:0003723 | 4 | 0.989 | 0.303 | 0.001 | 0.036 | 25 | 23 |
| zinc ion binding | molecular\_function | GO:0008270 | 6 | -0.880 | 0.266 | 0.001 | 0.036 | 25 | 24 |
| phosphorylation | biological\_process | GO:0016310 | 5 | 2.897 | 0.888 | 0.001 | 0.036 | 25 | 13 |
| organonitrogen compound catabolic process | biological\_process | GO:1901565 | 4 | -2.388 | 0.721 | 0.001 | 0.036 | 25 | 12 |
| endopeptidase activity | molecular\_function | GO:0004175 | 4 | -0.995 | 0.309 | 0.001 | 0.037 | 25 | 21 |
| pyrimidine-containing compound biosynthetic process | biological\_process | GO:0072528 | 5 | -5.505 | 1.711 | 0.001 | 0.037 | 25 | 7 |
| DNA recombination | biological\_process | GO:0006310 | 7 | -2.130 | 0.667 | 0.001 | 0.037 | 25 | 12 |
| oxidoreductase activity | molecular\_function | GO:0016491 | 2 | 2.541 | 0.801 | 0.002 | 0.037 | 25 | 13 |
| carbohydrate metabolic process | biological\_process | GO:0005975 | 3 | 2.245 | 0.717 | 0.002 | 0.039 | 25 | 15 |
| catalytic activity, acting on RNA | molecular\_function | GO:0140098 | 2 | -0.546 | 0.174 | 0.002 | 0.039 | 25 | 25 |
| pyrophosphatase activity | molecular\_function | GO:0016462 | 5 | -0.326 | 0.107 | 0.002 | 0.048 | 25 | 25 |
| organic cyclic compound binding | molecular\_function | GO:0097159 | 2 | 0.443 | 0.145 | 0.002 | 0.048 | 25 | 25 |
| hydrolase activity, acting on acid anhydrides | molecular\_function | GO:0016817 | 3 | -0.323 | 0.107 | 0.003 | 0.052 | 25 | 25 |

Table 5 Log2 median ratio counts of top taxa associated with COVID-19 (n= 29) compared to community acquired pneumonia (n=25) and uninfected (n=32) cohorts. Comparisons were conducted using Wilcoxon rank sum test and adjusted for multiple test comparisons using the Benajmini Hochberg correction method.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| treatment 1 | treatment 2 | log2 median ratio | median diff | mean diff | p value | q value | taxon name |
| COVID-19 | Community acquired pneumonia | -2.0539 | -0.0728 | -0.1332 | 0.0067 | 0.0224 | Paracoccus |
| COVID-19 | Community acquired pneumonia | -2.0539 | -0.0728 | -0.1332 | 0.0067 | 0.0224 | Sphingobium |
| COVID-19 | Community acquired pneumonia | -2.0539 | -0.0728 | -0.1332 | 0.0067 | 0.0224 | Sphingopyxis |
| COVID-19 | Community acquired pneumonia | -2.0539 | -0.0728 | -0.1332 | 0.0067 | 0.0224 | Sphingomonas |
| COVID-19 | Uninfected | -4.3829 | -0.4587 | -0.3767 | 0.0000 | 0.0000 | Paracoccus |
| COVID-19 | Uninfected | -4.3829 | -0.4587 | -0.3767 | 0.0000 | 0.0000 | Sphingobium |
| COVID-19 | Uninfected | -4.3829 | -0.4587 | -0.3767 | 0.0000 | 0.0000 | Sphingopyxis |
| COVID-19 | Uninfected | -4.3829 | -0.4587 | -0.3767 | 0.0000 | 0.0000 | Sphingomonas |
| COVID-19 | Uninfected | -5.1294 | -0.4164 | -0.3356 | 0.0000 | 0.0000 | Bradyrhizobium |
| COVID-19 | Uninfected | -5.1294 | -0.4164 | -0.3356 | 0.0000 | 0.0000 | Methylobacterium |

Table 6 Log2 median ratio counts of taxa associated with COVID-19 mortality when comparing deceased (n=10) versus survived (n=15). Comparisons were conducted using Wilcoxon rank sum test and adjusted for multiple test comparisons using the Benajmini Hochberg correction method.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| log2 median ratio | Median diff | Mean diff | p value | q value | Taxonomy |
| 2.25 | 0.361 | 0.371 | 0.00017 | 0.00691 | *Comamonadaceae* |
| 5.21 | 0.405 | 0.377 | 0.00017 | 0.00691 | *Variovorax* |
| 2.97 | 0.002 | 0.002 | 0.00353 | 0.074 | *Vibrionales* |
| 2.97 | 0.002 | 0.002 | 0.00353 | 0.074 | *Vibrionaceae* |
| 3.8 | 0.002 | 0.00181 | 0.00492 | 0.0827 | *Vibrio* |
| 1.84 | 0.0549 | 0.13 | 0.0137 | 0.124 | *Bacilli* |
| 2.24 | 0.403 | 0.297 | 0.0163 | 0.124 | *Burkholderiales* |
| 3.16 | 0.002 | 0.002 | 0.0157 | 0.124 | *Alteromonadales* |
| 3.61 | 0.004 | 0.004 | 0.0156 | 0.124 | *Yersiniaceae* |
| 2.1 | 0.005 | 0.00435 | 0.0156 | 0.124 | *Salmonella* |
| 1.77 | 0.011 | 0.064 | 0.0475 | 0.274 | *Streptococcaceae* |
| 2.29 | 0.425 | 0.296 | 0.0264 | 0.185 | *Betaproteobacteria* |
| -5.13 | -0.103 | -0.104 | 0.0308 | 0.199 | *Bacteroidia* |
| -5.18 | -0.099 | -0.102 | 0.00962 | 0.124 | *Bacteroidales* |



Figure 1. Heatmap with notable microbially derived gene ontology functional annotations associated with COVID-19 (n=32) as compared to community acquired pneumonia (n=29) & uninfected (n=25) cohorts. Rows are sorted by depth 1 parents and columns are clustered by Euclidean distance using ward D2 clustering. Comparisons were conducted using MaAsLin2, controlling for publication and patient ID with Benjamini Hochberg multiple test comparison (q<0.05)

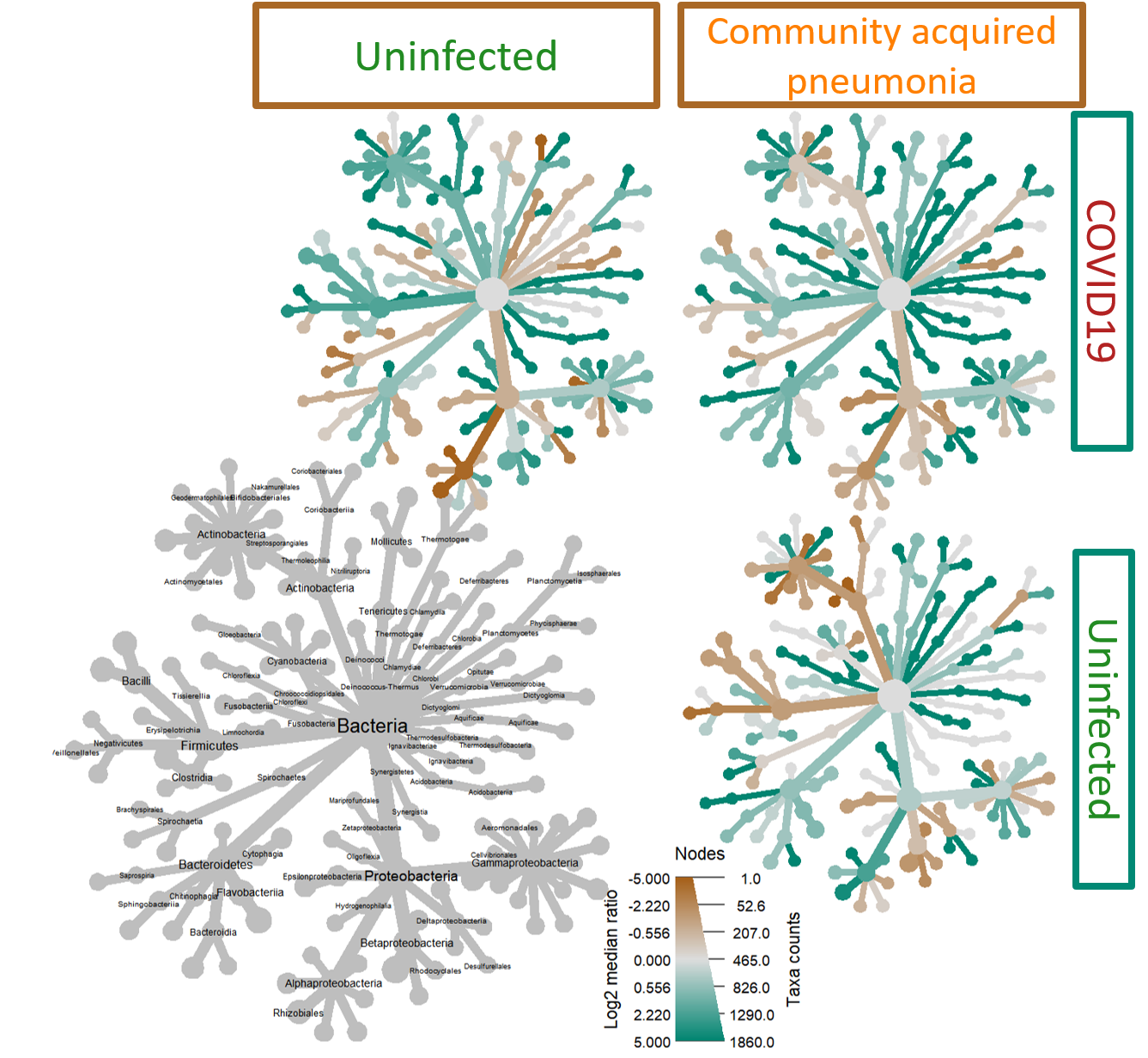


Figure 2. Heat tree matrix visualizing distinct COVID-19 vs Uninfected & viral pneumonia taxonomic profiles with significant changes in log2 median ration of several species belonging to the genus *Sphingomonas* when compared to both the uninfected (top left) and community acquired pneumonia cohorts (top right).



Figure 3 Heatmap of Significantly different gene ontology terms associated with COVID-19 mortality comparing deceased (n=10) versus survived (n=15). Rows are sorted by Depth1 parents and Columns are clustered by Euclidean distance using ward D2 clustering. Comparisons were conducted using MaAsLin2, controlling for patient ID with Benjamini Hochberg multiple test comparison (q<0.05)



Figure 4 Heat tree demonstrating the survivors of COVID-19 moderate to severe disease had a significantly unique BALF metatransciptome profile with notable decreases in the log2 median ratios in the Family *Comamonadaceae,* genus *Variovora,* and significant increases in the log2 median ratios of order *Bacteroidia* and class *Bacteroidales.*